Linear Regression Subjective Questions

1. What are the assumptions of linear regression regarding residuals?

   The four assumptions of linear regression regarding residuals are:
   - Linearity of residuals i.e. Drawing X and Y on a scatter plot should follows a linear pattern (i.e. not a curvilinear pattern) that shows that linearity assumption is met.
   - Independence of residuals i.e. no visible pattern should be present.
   - Normal distribution of residuals i.e. Error terms are normally distributed with mean zero.
   - Constant variance (homoscedasticity) of residuals i.e. the variance should not follow any pattern as the error values changes. A scatterplot of residuals versus predicted values is good way to check for homoscedasticity. There should be no clear pattern in the distribution. If there is a cone-shaped pattern, the data is heteroscedastic.

2. What is the coefficient of correlation and the coefficient of determination?

   Coefficient of correlation is the degree of relationship between two variables. It varies between -1 and 1. 1 indicates that the two variables are positively correlated i.e. they rise and fall together. -1 means that the two variables are negatively correlated i.e. if one goes up and other goes down. If both variables are not correlated then the correlation value will be 0.

   Coefficient of Determination is the square of Coefficient of Correlation. R square or Coefficient of Determination shows percentage variation on dependent variable which is explained by all independent variables together. It is always between 0 and 1. An R^2 of 0 means that the dependent variable cannot be predicted from the independent variable. An R^2 of 1 means the dependent variable can be predicted without error from the independent variable.

   Example: An R^2 of 0.10 means that 10 percent of the variance in *Y* is predictable from *X*

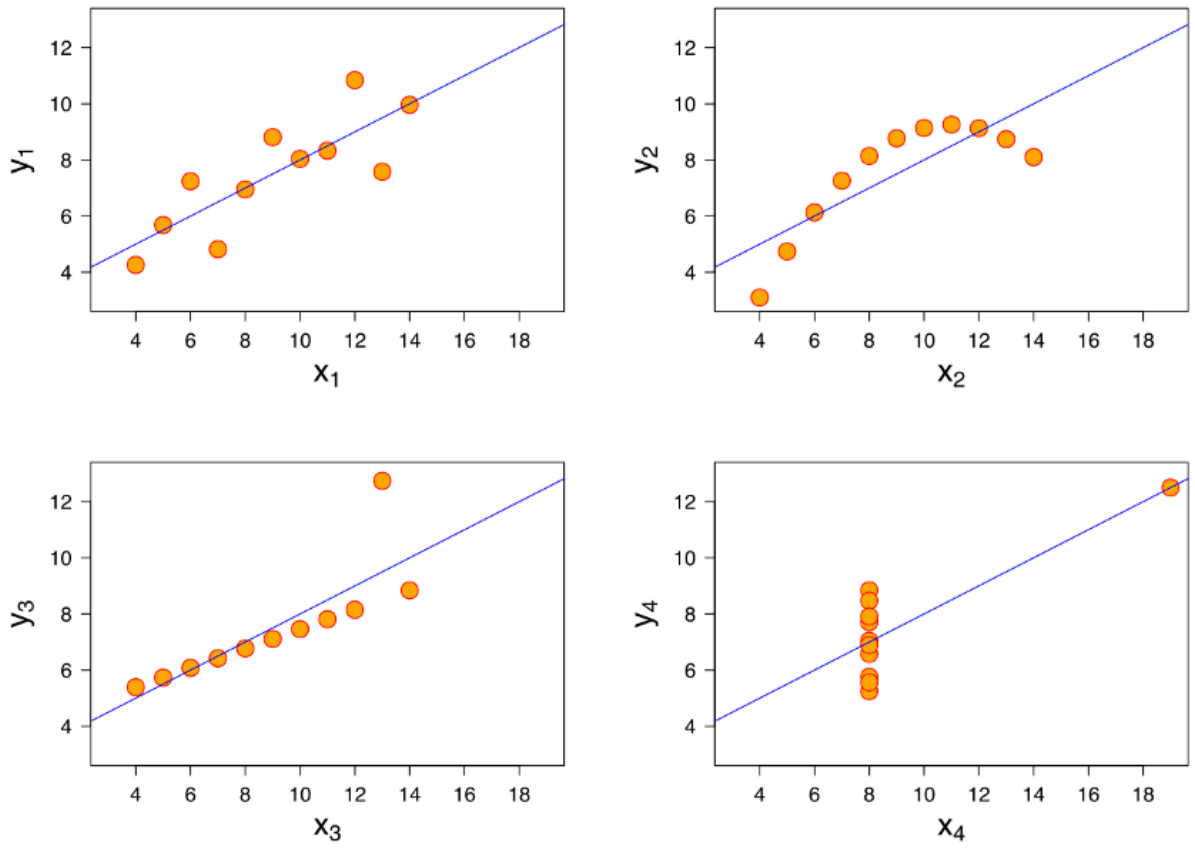
3. Explain the Anscombe's quartet in detail.

   Anscombe's quartet comprises of four datasets. Each dataset contains eleven (x,y) pairs. The important thing about these datasets is that they all share the same descriptive statistics but when graphed, each one tells a different story irrespective of their similar summary statistics.

   The summary statistics shows below three similarities for each dataset when drawn on x and y:
   - Mean of x is 9 and mean of y is 7.50
   - Variance of x is 11 and variance of y is 4.13
   - The correlation coefficient between x and y is 0.816

	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	8	5,56
	7	4,82	7	7,26	7	6,42	8	7,91
	5	5,68	5	4,74	5	5,73	8	6,89
SUM	99,00	82,51	99,00	82,51	99,00	82,50	99,00	82,51
AVG	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
STDEV	3,32	2,03	3,32	2,03	3,32	2,03	3,32	2,03

If we plot these four datasets on x-y plane, we can observe that they show the same regression lines as well but each dataset is telling a different story:



- Dataset I appear to have clean and well-fitting linear models.
- Dataset II is not distributed normally.
- Dataset III is linearly distributed, but the calculated regression is thrown off by an outlier.
- Dataset IV shows that one outlier is enough to produce a high correlation coefficient.

4. What is Pearson's R?

   Pearson's R or Pearson's correlation coefficient is the test statistics that measures the statistical relationship between two continuous variables. It is based on the method of covariance and gives information about the magnitude as well as the direction of the relationship (association). It can range from -1 to 1. An R of -1 indicates a perfect negative linear relationship between variables, an R of 0 indicates no linear relationship between variables, and an R of 1 indicates a perfect positive linear relationship between variables.

5. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

   Scaling is a method to bring all features to the same level of magnitudes i.e. to normalize the range of independent variables.

   Suppose if dataset contains features that are highly varying in magnitudes, units and range. If not taken care, these algorithms only take magnitude of features neglecting the units as most of the machine learning algorithms use Euclidean distance between two data points in their computations. E.g. If we take two quantity like 5kg and 5000gms. The features with high magnitudes i.e. 5000gms will weigh a lot more in the distance calculations than features with low magnitudes (i.e. 5kg) instead both are same.

   Normalization (Min-Max Scaling) typically means rescaling values into a range between 0 and 1. This might be useful in some cases where all parameters need to have the same positive scale. However, the outliers from the data set are lost.

$$z = \frac{x - \min(x)}{\max(x) - \min(x)}$$

   Standardization typically means rescaling values such that their mean comes out to be 0 with a standard deviation of 1.

$$z = \frac{x - \mu}{\sigma}$$

   where x is the data point, μ is mean of all data points and σ is standard deviation.

6. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

   This can be explained by VIF formula:

   $$VIF = 1/(1-R^2)$$

   i.e. if value of R^2 is 1, then denominator in above formula becomes 0 which overall return infinite (1/0).

   If R^2 of a variable is 1 then there exists another variable(s) which is very highly correlated to it and this will lead to multicollinearity. This should be taken care during model building.